Analysis of the second-hand housing price data of Loxton Town –Based on the linear regression model

Contents

1. Introducion **2**

2. Methodology **2**

2.1 Data . . . 2

2.2 Model . . . 5

3. Results **6**

4. Discussion **8**

4.1 Summary . . . 8

4.2 Conclusions . . . 8

4.3 Weakness & Next Steps . . . 9

References **10**

Appendix **10**

Preliminary modeling . . . 10

Eliminate Outliers . . . 11

Modeling Again (with or without stepwise regression) . . . 11

Heteroscedasticity Test . . . 12

Abstract: In recent years, as the price of first-hand housing has risen, the second-hand housing market has developed rapidly, and more and more people are willing and inclined to buy second-hand housing. Therefore, the study between second-hand housing prices and their influencing factors is of significant significance to buyers and the government. Based on the second-hand housing market data in Loxton, Australia from 2015 to 2020, this paper uses descriptive statistics, multiple linear regression models and stepwise regression methods to study the factors that affect the second-hand housing prices in this town, and proposes corresponding recommendations to buyers and the government.

Keywords: Loxton; Second-Hand Housing; Multiple Linear Regression;

1. Introducion

Housing price is an important index to measure people's livelihood and happiness index.

With the rapid development of economy and science and technology, the rapid rise of housing prices in central cities makes more and more young people choose to escape, which directly leads to a significant increase in housing prices in surrounding cities and towns. In addition, the development of the towns also has a positive role in promoting their housing prices to a certain extent.

Unfortunately, most of the studies are more concerned about the change of housing prices in central cities, and the impact on housing prices in towns is missing. For example, ang Li(2020) conducted a correlation study on women's fertility intention and housing price in all central cities of Australia, and found that in cities with high housing prices, fertility intentions could not be converted into housing relocation intentions, but could be converted in cities with lower housing prices. Morteza moallemi et al. (2020) studied the impact of domestic migration on housing prices in Australia, and found that immigration can promote housing prices.

In addition, the above literature is on the price of the first-hand housing market. However, no matter in the central cities or towns, with the gradual increase of the first-hand housing price, the second-hand housing market has become more and more prosperous, and attracted a lot of young people who have no higher economic ability and debt ability. Therefore, the research on the influencing factors of housing price in the second-hand housing market has a positive effect on the government's macro-control and policy guidance.

Therefore, this article takes the small town of Loxton, Australia as the research object, and the data span is 2015 to 2020. A multiple linear regression model is established for the price of second-hand housing in Loxton and its potential influencing factors to explore the impact of certain explanatory variables on housing prices.

2. Methodology

2.1 Data

The data in this article comes from the second-hand housing market data of the National Housing Authority of Australia, including part of the second-hand housing sales data in Loxton from 2015 to 2020. Among them, the explanatory variables include AGE, SHOPS, CRIME, TOWN, STORIES, OCEAN, POOL, SELLER, SIZE, SUBURB, TENNIS and SOLD and the dependent variable is PRICE.

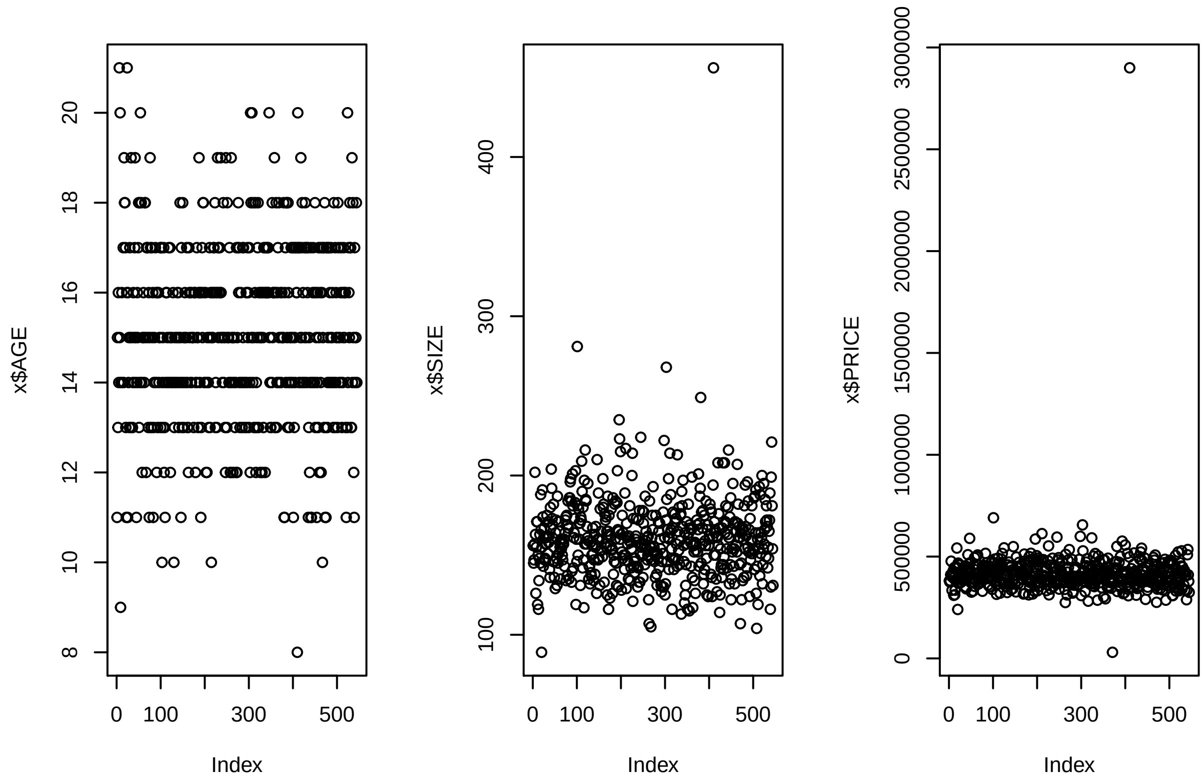


Fig 1. Scatter Plot

In Figure 1, we have drawn a scatter plot of three continuous variables of AGE, SIZE and PRICE. It can be seen from the figure that in the 545 observed samples, the value range of AGE is basically concentrated in 12 to 18 years, indicating that the residence life of this part of the second-hand housing samples is relatively long and the values are relatively concentrated. In addition, the value of the SIZE variable is more concentrated than that of AGE, which fluctuates from 100 to 220 square meters, and there is an abnormal value greater than 400 square meters. Finally, the concentration of PRICE variables is similar to SIZE, and there are also outliers.

Among them, OBS represents the serial number of each transaction and price is the dependent variable. After removing these two variables, there are 12 variables left. According to the different types of values, we preliminarily divide them into continuous variables and discrete variables. The continuous variables include AGE, CRIME, TOWN, STORIES, SIZE and SOLD, and the discrete variables include SHOPS, OCEAN, SELLER, SUBURB, TENNIS and POOL. We first give the descriptive statistics of several continuous variables and the category statistics of discrete variables. Table 1 and Figure 2 implies the former.

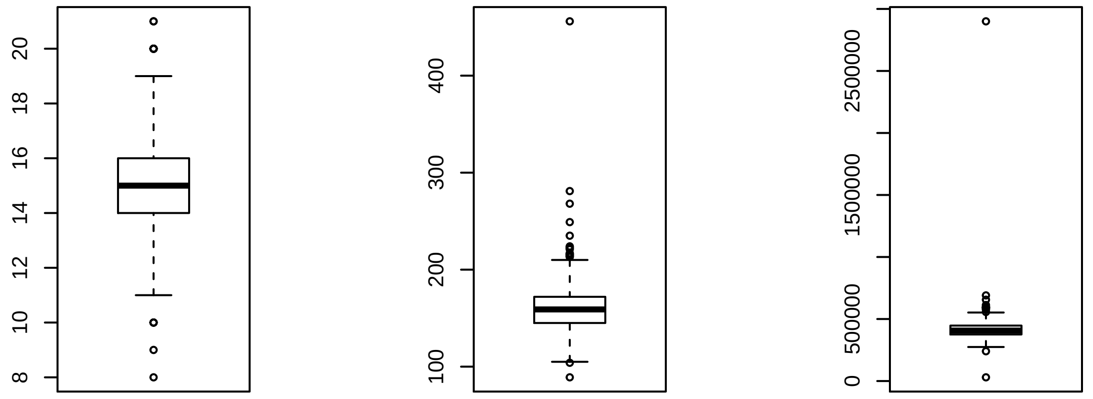
Table 1. Descriptive Statistics of Continuous Variables and Price

Statistics	AGE	CRIME	TOWN	STORIES	SIZE	PRICE
Min	8.00	1.00	30.00	1.00	89.00	30000
Max	21.00	3.00	60.00	4.00	456.00	2900000
Mode	14.00	3.00	60.00	1.00	157.00	NA
Mean	15.03	2.50	52.57	1.16	160.50	41645.56

Statistics	AGE	CRIME	TOWN	STORIES	SIZE	PRICE
Median	15.00	3.00	60.00	1.00	159.00	40656.50
Variance	4.11	0.75	168.02	0.18	719.15	1.52*1010
Kurtosis	0.13	-0.63	-0.63	8.74	27.43	303.39
Skewness	-0.02	-1.17	-1.17	2.84	2.89	15.07

`Note: Two decimal places are reserved in the table`

It can be seen from the Table 1 that the minimum value of AGE variable is 8 years and the maximum value is 21 years, which indicates that the longest service life of houses sold in this town is 21 years. And we can conclude that the age of house is mostly concentarted on 15 years because of the difference between the mode, mean and median of AGE is very small while its skewness close to 0. On the other hand, the variance of PRICE variable reaches 1.52*1010, which indicates that the sales prices of houses in this town vary greatly and there are many factors would affecting the prices. For several other continuous variables, such as CRIME and STORIES. CRIME indicates the crime rate in the area where the house is located. The maximum value is only 3, which indicating that the town has a better security management for the house. STORIES represents the number of accidents in the house and its average, median and other statistics are also very small which also implies that the town's houses are safe and suitable for living. As for the variables TOWN and SIZE, TOWN represents the distance from the house to the town centre. The closer the distance, the greater the commercial value of the house. The average, median, and mode of TOWN are similarly close, and the skewness is also small, indicating the value of Town clusters around 50~60. SIZE represents the area of the house. From Table 1, it can be seen that the value range of this variable is large and the fluctuation is also large. In addition to summary statistics, we are also concerned about whether there are outliers in the data set. Therefore, Figure 2 shows box plots of variables AGE, SIZE and PRICE with large fluctuations.



Box plot of AGE

Box plot of SIZE

Box plot of PRICE

Fig 2. Box Plot

Of course, in addition to the above descriptive statistics of continuous variables, we also report some summary statistics of discrete variables. SHOPS indicates whether the house is close to the shopping area, if it is close, then its value is 1, 0 otherwise. But in the data set, we found that the value of this variable includes 2 and 3. Therefore, in order to fully consider the modeling, we adopt two treatment methods. The first is to delete the variable directly and the second is to take the variable as a Multi-valued dummy variable, not just 0-1 value. The remaining dummy variables are all 0-1 values, we will not describe too much. At last, Table 2 shows the correlation coefficient between the CRIME, TOWN and SUBURB. It can be clearly seen that the three variables have multicollinearity. Therefore, before modeling, we choose to keep only the CRIME variable.

Table 2. Correlation coefficient between the CRIME, Town and SUBURB

Corr	CRIME	TOWN	SUBURB
CRIME	1.00	1.00	-1.00
TOWN	1.00	1.00	-1.00
SUBURB	-1.00	-1.00	1.00

Note: Two decimal places are reserved in the table

2.2 Model

Since the data is cross-sectional data, we use a multiple linear regression model based on least squares estimation (OLS) for modeling and analysis. The estimated equation of the multiple linear regression model is

$$Y = \beta_0 + \beta_1 * X_1 + ... + \beta_k * X_k.$$

Among them, β_0 is the intercept term of the model, and $\beta_i (i > 1)$ is the coefficient of the explanatory variable, also called the regression coefficient, which reflects the degree of influence of the explanatory variable on the dependent variable Y. Least squares estimation (OLS) is an estimation method based on the criterion of mean square error minimization. The coefficients obtained by OLS have the characteristics of best linear unbiased estimation (BLUE).

Below, we give a simple description of the least square estimation of the multiple linear regression model. The assumptions of the multiple linear regression model mainly include two. First, there is a multiple linear relationship between the independent variable and the dependent variable, the dependent variable y can be completely linearly explained by $x_1, x_2, ..., x_k$. Second, the part that cannot be explained is pure unobservable error. Then, the purpose of our modeling is to estimate the parameter $\beta_0, \beta_1, ..., \beta_k$. Based on the expression of the minimum mean square error, we can obtain the estimation equation of the parameters of the multiple linear regression model by calculate, derivative and so on as (matrix expression)

$$\beta = (X\text{'}X)^{(-1)} X\text{'}Y.$$

Then, we will test the model at two levels. The first part is to test the significance of a single parameter of the model. That is, the null hypothesis H0: $\beta_i = 0$, which is tested by constructing t statistic. The second part is to test the overall significance of the model which called F test, that is, the null hypothesis H0: $\beta_0 = \beta_1 = ... = \beta_k = 0$.

In addition, the multiple linear regression model includes two tests, one is the t test on the explanatory variable coefficients, and the other is the F test on the model. These two test choices variables that can have a significant impact on the dependent variable when modeling. On the other hand, we can also use

AIC, BIC and other information criteria to select variables, which is also the basis for forward regression, backward regression and stepwise regression. From the descriptive statistics of the data in Section 2.1, there are three variables in the data that have serious multicollinearity and outliers. At the same time, considering the heteroscedasticity of the model, we first conduct a preliminary modeling of the data. Therefore, we first model the data, and then use the stepwise regression method to screen variables, and judge the model's quality based on the t-test results and the R-square of the model.

Of course, since there are dummy variables with values of 0 or 1 in the data, we need to fit a multiple linear regression model with dummy variables. At this time the form of the model becomes

$$Y = \beta_0 + \beta_1 * X_1 + ... + \beta_k * X_k + \alpha_1 * D_1 + ... + \alpha_j * D_j,$$

where D_i means the dummy variables.

Dummy variable regression is a very basic but very practical model in the field of econometrics. When the explanatory variables are dummy variables, we can still use traditional methods such as least squares(OLS) and maximum likelihood estimation(MLE) to fit the model. But if the dependent variable becomes a dummy variable, the model obtained at this time is called logistic regression, probit regression, etc.

3. Results

Based on the results of Section 2, we first establish a multiple linear regression model and test and diagnose the model. The dependent variable of the model is PRICE and the remaining 12 variables are used as explanatory variables to model the dependent variables. Table 3 shows the results of modeling PRICE and 12 explanatory variables. It can be seen from Table 3 that the effect of modeling on the 12 explanatory variables is very poor. The R-square is only 0.66 and the significance test of multiple variables fails. We can preliminary judge that model is affected by outliers and multicollinearity.

Table 3. Modeling result 1

Variables	Estimate	P-value
Intercept	56295.6000	0.1212
AGE	-2755.5000	$0.0748^{.}$
CRIME	-14959.7000	$3.8100 * 10^{-5***}$
STORIES	42098.4000	$2.6500 * 10^{-8***}$
SIZE	3203.1000	$< 2 * 10^{-16***}$
OCEAN	-75969.0000	$< 2 * 10^{-16***}$
POOL	-26427.3000	0.0037^{**}
SELLER	-18774.5000	0.0123^{*}
TENNIS	-25400.3000	0.0058^{**}
Multiple R-squared: 0.6605	Adjusted R-squared: 0.6555	F-statistic: 130.4

Note: ***, **, * and . represent the significance level of 0.001, 0.05, 0.01 and 0.1 respectively.And two decimal places are reserved in the table.

Next, combining the results of Figure 1 and Table 3, we further tested the data for outliers and exclude outliers that AGE greater than 20 or less than 10, SIZE greater than 250 or less than 100 and PRICE greater than 1000000 or less than 200000. In addition, in order to eliminate the effects of multicollinearity and potential heteroscedasticity, we use stepwise regression to model the data set and delete variables that have no significant impact on PRICE. The result without stepwise regression are given in Table 4 and the result with stepwise regression are given in Table 5.

Table 4. Modeling result 2

Variables	Estimate	P-value
Intercept	121845.9300	$< 2 * 10^{-16***}$
AGE	1230.4500	$0.0144^{*}*$
TOWN	-1027.3500	$< 2 * 10^{-16***}$
SIZE	1941.0100	$< 2 * 10^{-16***}$
OCEAN	67694.5900	$< 2 * 10^{-16***}$
POOL	7203.1200	0.0112^{*}
SELLER	10347.7200	$9.76 * 10^{-6***}$
TENNIS	17893.2900	$9.72 * 10^{-10***}$
Multiple R-squared: 0.8540	Adjusted R-squared: 0.8500	F-statistic: 218

Note: ***, **, * and . represent the significance level of 0.001, 0.05, 0.01 and 0.1 respectively.And two decimal places are reserved in the table.

Table 5. Modeling result 3

Variables	Estimate	P-value
Intercept	118457.1100	$< 2^{*}\ 10^{-16***}$
AGE	1353.6100	0.0064^{**}
TOWN	-1019.2400	$< 2 * 10^{-16***}$
SIZE	1949.0900	$< 2 * 10^{-16***}$
OCEAN	67660.3900	$< 2 * 10^{-16***}$
POOL	7009.5600	0.0138^{*}
SELLER	10348.7700	$9.80 * 10^{-6***}$
TENNIS	17341.9700	$2.75 * 10^{-9***}$
Multiple R-squared: 0.8511	Adjusted R-squared: 0.8491	F-statistic: 431.9

Note: ***, **, * and . represent the significance level of 0.001, 0.05, 0.01 and 0.1 respectively.And two decimal places are reserved in the table.

Because of the obvious multicollinearity of CRIME, TOWN and SUBURB, we finally chose the stepwise regression model with R-squared = 0.8511. At last, we test the model for heteroscedasticity furthermore. The heteroscedasticity test of the model is used to test whether there is heteroscedasticity in the model. The GQ-test is commonly used and the P value compute by GQ test in this model is 0.146, which is much greater than the significance level of 0.05, indicating that the null hypothesis is accepted and there is no heteroscedasticity in the model.

In summary, the final model we get is

$$PRICE = 118457.11 + 1353.61 * AGE - 1019.24 * TOWN + 1949.09 * SIZE + 67660.39$$

$$*OCEAN + 7009.56 * POOL + 10348.77 * SELLER + 17341.97 * TENNIS$$

4. Discussion

4.1 Summary

Based on the data of second-hand housing prices and their influencing factors in loxton from 2015 to 2020, this paper establishes a multiple linear regression model with dummy variables on the basis of summary statistics, selects variables through stepwise regression and other methods, and finally determines seven variables such as AGE and TOWN have a significant impact on the housing prices of second-hand houses in this town.

The model we built is carried out on the basis of eliminating outliers and multicollinearity and all parameters pass the t-test and the model also passes the F-test while R-squared=0.8511. At the end of the modeling, we also performed a heteroscedasticity test and all the above results show that our model is suitable.

4.2 Conclusions

As we can see, the continuous variables that have a significant impact on the housing price of loxton second-hand houses are AGE and SIZE. Their coefficients represent the meaning that every increase in house age by 1 year or every increase in house area by 1 square meter will increase the house price AUD 1353.61 or AUD 1949.09. Secondly, among the remaining 5 variables that have a significant impact on housing prices, TOWN is worth noting, because although its value is an integer greater than 1, its value only contains 30 and 60. According to the definition of dummy variables , we can still regard TOWN as a dummy variable. The coefficient of TOWN means that increasing the value of TOWN from 30 to 60 will reduce house prices by AUD 1019.24. In addition, we should also note that OCEAN has the largest impact on house prices among the various variables, which means that a house with and without sea views will cause the price difference of AUD 67660.39. Finally, when the values of all variables are 0, the house price is AUD 118,457.11.

In addition, we can also make a simple prediction and solve the 95% confidence interval of the predicted value. Suppose there is a customer named Jane who wants to sell her house and we are aimed at predicting the price of her house. Her house has two stories, is 192 square metres large, is not near a shopping precinct and is 10 km from the town centre. She estimates that the house is about 10 years old and in a low crime area according to her experiences. Jane inherited the house from her uncle and is therefore unsure when it was last sold. Moreover, her house has tennis courts and sea views. As for whether the house possesses a pool, we have no assumption to judge it and regard it does not have a pool. Based on this model, we can predict Jane's house selling price. Hence, we can assign values to variables in the model. AGE=10, TOWN=10, SIZE=192, OCEAN=1, POOL=0, TENNIS=1.The value of SELLER cannot be determined, so we will discuss it in different situations. The calculation formula for the confidence interval of the multiple regression model point estimate is

$$[\hat{Y}_F - \hat{\sigma}\sqrt{1 + X_F(X'X)^{-1}X_F'} * t_{\alpha/2}(n-k-1), \hat{Y}_F + \hat{\sigma}\sqrt{1 + X_F(X'X)^{-1}X_F'} * t_{\alpha/2}(n-k-1)]$$

If we choice W&M, then $SELLER = 0$ and

$$PRICE = 118457.11 + 1353.61 * 10 - 1019.24 * 10 + 1949.09 * 192 + 67660.39$$

$$*1 + 7009.56 * 0 + 10348.77 * 0 + 17341.97 * 1 = 581028.45.$$

The standard deviation $\hat{\sigma}$ of residuals is 22212.48 and hence the 95% confidence interval for this sale price is

$$[581028.45 - 22212.48 * 1.0274 * 1.9644, 581028.45 + 22212.48 * 1.0274 * 1.9644] = [536198.68, 625858.22].$$

And W&M charges a commission of 5%, hence, Jane will pay a commission of AUD 29051.42. If we choice A&B, then $SELLER = 1$ and

$$PRICE = 118457.11 + 1353.61 * 10 - 1019.24 * 10 + 1949.09 * 192 + 67660.39$$

$$*1 + 7009.56 * 0 + 10348.77 * 1 + 17341.97 * 1 = 591377.22.$$

The 95% confidence interval for this sale price is

$$[591377.22 - 22212.48 * 1.0311 * 1.9644, 591377.22 + 22212.48 * 1.0311 * 1.9644] = [546386.00, 636368.44].$$

And A&B charges a commission of 10%, hence, Jane will pay a commission of AUD 59137.72. Comparing the above two situations, we suggest Jane use W&M to sell her house. The above is a simple example of model prediction. We have given two point estimates and interval estimates when the value of SELLER is uncertain.

4.3 Weakness & Next Steps

In section 3, we first perform a preliminary modeling of all the data and then based on the results, we find that there are outliers and multicollinearity in the model, so we deal with these two issues and use the stepwise regression method to modeling again. We finally get the model with R-square=0.8511 at this time and point out the following problems in the modeling process and the results.

Problem 1: Model has the misspecification influence which comes from two parts. The first part is that we only deleted outliers that exceed 90% quantile and less than 10% quantile, but this does not mean that the samples in 80%~90% and 10%~20% are not outliers, which require further modeling and verification and will have a small impact on the results in section 3.

Problem 2: There are some problems in the original data. First, the description of the data set does not match the actual value, resulting in ambiguity in the modeling process such as SHOPS contains 0,1,2,3, while in data definitions, only mentioned 0 and 1. Second, CRIME, TOWN, SUBURB have serious collinearity which may be caused by incorrect data records. At last, in our model, the variable SOLD is not significant, which may be due to the fact that the data set samples are relatively small compared to the variables leads to simple multiple linear regression cannot fit well.

Problem 3: There are too many dummy variables in the data set, so that the model cannot learn enough information from fewer samples. In addition, there is multicollinearity between the two variables of CRIME and TOWN and their values are not continuous, which makes it possible to treat them as dummy variables during modeling.In the subsequent modeling process, the first step is to modify and expand the data set and obtain enough samples while ensuring the accuracy of the data.

Problem 4: The coefficient of AGE is positive, which is contrary to the actual situation. Generally speaking, the longer the residence life of a house, the lower the value of the house and the sales potential. But from the results of the model, the greater the residence period, the house price will increase. The reasons for this result may include two aspects. The first is comes from the data . The minimum value of the AGE is 8 years and the maximum value is 21 years. The span is not large enough to fully reflect the overall impact of house age on housing prices. The second is the subjective thinking of buyers. For potential buyers, the shortest residence life is 8 years, which leads to a decrease in the influence of the residence life on the purchase intention at this time. That is to say, in the eyes of buyers, there is little difference between 8 years and 21 years, and longer length of residence means that houses are more popular.

Based on the above problems, our future work can be carried out from the following three aspects. First of all, we need to re-collect and search data, which would expand the number of samples and variables as much as possible, and ensure the accuracy and authenticity of the data. Second, we need to consider more data preprocessing methods. In the section 2 of this article, we only performed descriptive statistics on the data. In future work, since OCEAN and POOL are obviously related, We can linearly combine these two variables in the preprocessing stage, and mine the information in the data in advance through prior knowledge. Finally, we need to combine richer and more complex modeling methods. When the sample size and the number of variables gradually expand, it is obvious that the multiple linear regression model cannot match the data. At this time, models that consider interaction effects or are based on some dimensionality reduction methods To filter variables can better train the model.

References

Chun-Chang Lee, Chih-Min Liang, Jian-Zheng Chen, et al. Effects of the housing price to income ratio on tenure choice in Taiwan: forecasting performance of the hierarchical generalized linear model and traditional binary logistic regression model. 2018, 33(4):675-694.

Md Abdullah Al-Masum, Chyi Lin Lee. Modelling housing prices and market fundamentals: evidence from the Sydney housing market. 2019, 12(4):746-762.

Wei Zheng, Xinyi Li, Nanxing Guan, et al. Correlation Analysis of Fiscal Revenue and Housing Sales Price Based on Multiple Linear Regression Model. 2020, 9:3-12.

Appendix

In this section, we give detailed modeling steps and code.

Preliminary modeling

```
lm5.6<-lm(x$PRICE~x$AGE+x$CRIME+x$STORIES+x$SIZE+x$OCEAN+x$POOL+x$SELLER+x$TENNIS+x$SOLD,data=x)
lm5.6_step<-step(lm5.6,direction="both",trace=0)
summary(lm5.6_step)
```

```
##
## Call:
## lm(formula = x$PRICE ~ x$AGE + x$CRIME + x$STORIES + x$SIZE +
##     x$OCEAN + x$POOL + x$SELLER + x$TENNIS, data = x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -341260  -28861    -106   30124 1307002
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -90275.5    32487.2  -2.779  0.00565 **
## x$AGE        -2755.5     1543.7  -1.785  0.07483 .
## x$CRIME     -14959.7     3601.7  -4.154 3.81e-05 ***
## x$STORIES    42098.4     7454.6   5.647 2.65e-08 ***
## x$SIZE        3203.1      119.7  26.765  < 2e-16 ***
## x$OCEAN      75969.0     8885.3   8.550  < 2e-16 ***
## x$POOL       26427.3     9050.6   2.920  0.00365 **
## x$SELLER     18774.5     7470.7   2.513  0.01226 *
## x$TENNIS     25400.3     9165.1   2.771  0.00578 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72370 on 536 degrees of freedom
## Multiple R-squared:  0.6605, Adjusted R-squared:  0.6555
## F-statistic: 130.4 on 8 and 536 DF,  p-value: < 2.2e-16
```

Eliminate Outliers

```
x<-subset(x, AGE<=20)
x<-subset(x, AGE>=10)
x<-subset(x, PRICE<=1000000)
x<-subset(x, PRICE>=200000)
x<-subset(x, SIZE<=250)
x<-subset(x, SIZE>=100)
```

Modeling Again (with or without stepwise regression)

```
lm5.7<-lm(x$PRICE~x$AGE+x$CRIME+x$STORIES+x$SIZE+x$OCEAN+x$POOL+x$SELLER+x$TENNIS+x$SOLD,data=x)
#lm5.7_step<-step(lm5.7,direction="both")
summary(lm5.7)
```

```
##
## Call:
## lm(formula = x$PRICE ~ x$AGE + x$CRIME + x$STORIES + x$SIZE +
##     x$OCEAN + x$POOL + x$SELLER + x$TENNIS + x$SOLD, data = x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -56261 -14708  -2689  14466  67983
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  894827.70 1311610.80   0.682  0.49539
## x$AGE          1338.76     495.39   2.702  0.00711 **
## x$CRIME      -15273.19    1116.98 -13.674  < 2e-16 ***
## x$STORIES      1654.85    2378.83   0.696  0.48695
## x$SIZE         1947.96      43.29  44.994  < 2e-16 ***
## x$OCEAN       67591.38    2771.53  24.388  < 2e-16 ***
## x$POOL         7019.33    2835.50   2.476  0.01362 *
## x$SELLER      10356.79    2322.08   4.460 1.00e-05 ***
## x$TENNIS      17267.01    2872.02   6.012 3.42e-09 ***
## x$SOLD         -393.16     650.09  -0.605  0.54559
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22380 on 527 degrees of freedom
## Multiple R-squared:  0.8513, Adjusted R-squared:  0.8488
## F-statistic: 335.3 on 9 and 527 DF,  p-value: < 2.2e-16
```

```
lm5.8<-lm(x$PRICE~x$AGE+x$CRIME+x$STORIES+x$SIZE+x$OCEAN+x$POOL+x$SELLER+x$TENNIS+x$SOLD,data=x)
lm5.8_step<-step(lm5.8,direction="both",trace=0)
summary(lm5.8_step)
```

```
##
## Call:
```

```
## lm(formula = x$PRICE ~ x$AGE + x$CRIME + x$SIZE + x$OCEAN + x$POOL + 
##     x$SELLER + x$TENNIS, data = x)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -56002 -14222  -2137  14320  67957 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 103168.47   10457.45   9.866  < 2e-16 ***
## x$AGE         1353.61     494.54   2.737  0.00641 ** 
## x$CRIME     -15288.64    1115.06 -13.711  < 2e-16 ***
## x$SIZE        1949.09      43.22  45.092  < 2e-16 ***
## x$OCEAN      67660.39    2767.48  24.448  < 2e-16 ***
## x$POOL        7009.56    2830.03   2.477  0.01357 *  
## x$SELLER     10348.77    2317.88   4.465 9.80e-06 ***
## x$TENNIS     17341.97    2866.74   6.049 2.75e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 22360 on 529 degrees of freedom
## Multiple R-squared:  0.8511, Adjusted R-squared:  0.8491 
## F-statistic: 431.9 on 7 and 529 DF,  p-value: < 2.2e-16
```

Heteroscedasticity Test

```r
library(lmtest)
```

```
## Loading required package: zoo
```

```
## 
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
## 
##     as.Date, as.Date.numeric
```

```r
gqtest(lm5.8_step)
```

```
## 
##  Goldfeld-Quandt test
## 
## data:  lm5.8_step
## GQ = 1.1397, df1 = 261, df2 = 260, p-value = 0.146
## alternative hypothesis: variance increases from segment 1 to 2
```